广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(4):483~485,492
Vol. 26, No. 4 November 2010

基于数据世系的数据质量评估系统框架设计*

The Framework of Data Quality Assessment Based on Data Provenance

郑 华
ZHENG Hua

(广西财经学院计算机与信息管理系,广西南宁 530003)
(Department of Computer and Information Managing, Guangxi University of Finance and Economics, Nanning, Guangxi, 530003, China)

摘要:为了从源头解决数据质量问题,实现数据的可溯源,设计一个基于数据世系的数据质量评估系统框架。该系统框架可以动态添加各种不同的评估方法,通过数据世系技术分析出数据的原始演化过程进而识别出其维度,再选择系统中合适的评估方法实现数据质量评估。

关键词:数据 质量评估 质量管理 数据世系 溯源
中图法分类号:TP315 **文献标识码**:A **文章编号**:1002-7378(2010)04-0483-03

Abstract: In order to solve data quality problems from the source and realize the traceability of data, the framework of data quality assessment based on data provenance is presented. The framework can dynamically add various different assessment functions, on which the evolution of data is analyzed; the dimensions are identified; the appropriate assessment functions are selected to realize data quality assessment.

Key words: data, quality assessment, quality management, data provenance, traceability

学术界针对数据质量的研究主要围绕数据质量提高技术和数据质量评估两个方面展开。数据质量评估是数据质量管理的源头问题,其核心在于如何具体地评估各个维度,但是一直以来学术界缺乏足够的重视。数据质量评估依赖于使用数据的个体,不同环境下不同人员的"使用的适合性"各不相同[1]。数据质量是相对的,不能独立于使用数据的消费者来评价数据质量。因此,识别数据质量维度成为有价值的研究工作。从已有的数据质量维度方案可以看出,由于数据质量依赖使用数据的个体,无法也没有必要建立一套广泛接受的数据质量维度。在现实中,一个可行的方案就是针对特定的环境识别出相应的数据质量维度。

目前数据质量评估的方法主要分成定性的策略和定量的策略两类[2,3]。对各个维度从定性的角度来分析其"好"或"坏",这是目前数据质量评估方法的主流。国内外学者主要从方法论角度研究数据质量评估。文献[4]总结了当前主流的包括 TDQM、DWQ、TIQM、AIMQ 等 13 种数据质量评估方法。杨清云等[5]设计一个六元组的数据质量评估模型,实现了通过量化的指标来对应用系统中的数据质量状况进行评估。这些评估方法虽然都不一样,但是都有其适用的范围,共同点就是主观和客观相结合来完成评估活动。

数据世系指的是数据的产生、并随时间推移而演化的整个过程的信息[6]。数据世系包含静态的源数据信息和动态的数据演化过程,其特点是侧重于描述各种不同应用中数据来源和演化过程的信息,包含了更加丰富的元数据信息。数据质量管理是数据世系的一个重要的应用领域,特别是当前的互联网、网格、云计算等各种数据密集型应用导致了更为复杂的数据质量问题。如果仅凭独立的数据集合难以轻易判定其质量高低,必须结合世系信息深入了

收稿日期:2010-09-21

作者简介:郑 华(1978-),男,副教授,主要从事网络管理信息系统研究。

*广西哲学社会科学"十一五"规划 2008 年度立项项目(编号:08FTQ001)资助。

解数据产生、演化的具体过程才能合理评估数据集合的质量。另一方面，目前企业信息化过程中的热点-数据集成，关注的是集成后的目标数据，而忽略了数据的来源演化过程。面对更加复杂的应用环境往往由于源头数据的质量问题难于实现集成目标。但是，当前数据质量评估多从方法论角度来切入，其对象是目标数据本身，没有考虑到数据演化的过程对质量的影响因素。因此本文提出一种新的思路，设计一个系统框架，动态添加各种不同的评估方法，通过数据世系技术分析出数据的原始演化过程进而识别出其维度，再选择系统中合适的评估方法实现数据质量评估。

1 框架的结构设计

我们利用面向服务的体系结构(SOA)思想，将数据质量评估系统架构进行抽象化，使每个服务变得清晰，最终实现系统可以以较低的成本进行重组，体现评估系统的快速可重构性，创建一个能适应各类不同环境下评估需求的框架结构。数据质量评估系统把应用系统的功能及业务流程逻辑封装成标准的服务，通过服务的描述、发布与发现机制实现服务间的调用与组合，最终实现数据质量评价的整个业务流程。系统的框架结构如图1所示。

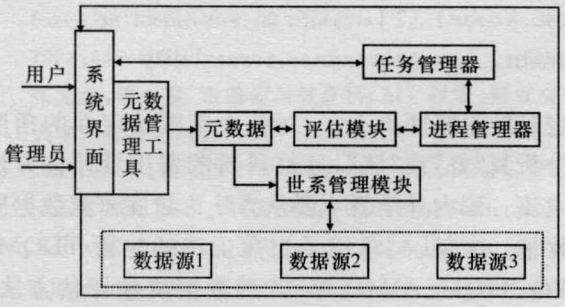

图1 数据质量评估系统框架结构

基于SOA的数据质量评估系统的核心为评估模块和世系管理模块。评估模块中各种评估组件以Web service的形式动态添加和设置以适应不同的评估要求。世系管理模块则负责世系数据的采集和查询功能。

通过采用SOA框架，可以最大程度地减少系统间的耦合，从而提高可重用性。通过研究相应的组件和应用来组合产生新的服务，并针对不同环境需求，灵活的构建应用程序和评估流程。

2 框架的世系管理流程设计

世系管理流程关注的是数据内容及其演化过程，其核心在于对数据的溯源。数据质量评估系统框架的世系管理业务流程由计划和组织、主数据一致性、记录溯源数据、提出溯源请求以及实施溯源5个子流程组成(图2)。

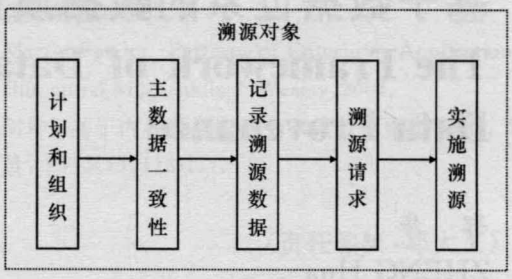

图2 世系管理流程

计划和组织子流程它实施世系管理的前提，主要执行者是数据管理的各个溯源参与方。该子流程主要功能是确定溯源数据的分配、采集、共享和保存方式，以及输入、内部流程与输出之间链接的管理方式。

主数据一致性子流程决定如何为对象、物理位置分配标识，以及对象间进行主数据交换的方式。主要包括为对象分配标识、为物理位置分配标识、交换主数据等步骤。

记录溯源数据子流程决定溯源对象标识的分配、应用和采集方式，以及整个世系管理流程中溯源数据的采集、共享和保存方式。

溯源请求子流程确定溯源请求的提出和响应方式，任何溯源参与方均可提出溯源请求。为实现溯源请求的目标，一个溯源请求可以引发整个流程中向上或向下若干步的后续溯源请求，使得“向上一步和向下一步”溯源有效。为了更快的采集信息，溯源请求可跳过一个步骤，与更上一个或更下一个溯源参与方取得联系。当需要进行溯源时，该子流程开始。溯源请求子流程包括提出溯源请求、接收溯源请求、回复溯源请求、接收溯源请求的答复等步骤。

实施溯源使数据质量评估模块能够根据实际要求，运用预设的溯源流程对溯源对象实施溯源。该子流程的结果是溯源参与方对溯源对象实施了源，并满足评估的要求。

3 关键技术

3.1 世系的语义模型

定义1 一个命名空间(name space)表示的是无限可数集，可以记作NS，所有的字符串集可以看作是一个命名空间。

定义2 一个命名值(name value)记为(name, value-list)，name∈NS，value-list是一个数据列表，

可以为空。

定义 3 识别码(identifier)可以看作是一个命名值。

定义 4 一个应用环境(application environment)可以记为(NS,CLK,ADDR,DICT),NS 就是一个命名空间,CLK 表示的是系统确认时间,ADDR 表示的是系统确认地址,DICTNS,是一个预先定义的字符集。用来表示数据的最小个体并提供某种程度的交互。

定义 5 一个标注(annotation)是一套命名值,每一个命名值可以看作是一个入口(entry),用于描述数据的一种扩展方式。

定义 6 一个静态实体记录(static entity record)可以记为(entity-id, entity-address, entity-type, annotation, snapshot-time, record-id),entity-id 可以看作是识别码(identifier), entity-address⊆ADDR, entity-type⊆DICT, annotation 参看前面的定义, snapshot-time⊆CLK, record-id⊆NS。这个对象捕捉的是某个时刻某个实体的某些方面信息。

定义 7 一个活动类型(activity type)可以记作(type-name, incoming-list, outgoing-list, annotation),type-name 就是一个识别码. incoming-list 是一列有序的命名值(该列中的每个命名值称作输入队列). outgoing-list 也是一列有序的命名值(该列中的每个命名值称作输出队列).

定义 8 一个活动记录(activity record)可以记作(activity-id, activity-type, activity-span, annotation, record-id), activity-span = (start, end), start and end⊆CLK。

基于以上的定义,我们给出世系实体及其关系的描述,一个世系实体可以看作是一个静态实体记录或一个行为记录,定义为(type, entity),type⊆{S,A},如果 type=S, 该世系实体就是一个静态实体记录,如果 type=A,该实体就是一个活动记录;一个世系关系就是两个世系实体间的关系,定义为(causal-entity, consequential-entity, role, annotation, relationship-id),causal-entity 和 consequential-entity 都是世系实体,role⊆DICT,relationship-id⊆NS。通过这个语义模型的定义,可以以一种灵活的方式去描述应用系统中所有静态和动态元素的世系信息,

3.2 世系的查询模型

基于 3.1 中世系的语义模型定义,我们开发出一种能够操作世系实体和关系的查询模型,主要包括两方面的内容。

3.2.1 世系实体的查询

由 3.1 的定义知道,世系实体有两种情况:静态实体记录和活动记录,一个世系实体是可以访问的,因此我们把世系实体的存储定义为一个可访问的对象:世系实体存储。要查询世系实体的信息主要通过定义对世系实体存储的操作器来完成。

3.2.2 世系关系的查询

从 3.1 的语义模型知道,世系关系其实就是两个世系实体间的关系,我们这里同样定义一个世系关系存储作为世系关系的存储对象,要查询世系实体关系的信息也是要通过定义对世系实体关系存储的操作器来完成。

整个查询模型由世系实体存储、世系关系存储和各种查询操作器构成。和一般数据查询不同的是,世系查询的结果不仅包括复杂的内容结构,还需要包括复杂的关系结构。限于篇幅,对操作器的定义不在此详细列出。

4 结束语

从已有的研究成果来看,通用的数据质量评估维度和评估方法是不可行的,需要针对不同的应用环境和个体分析其维度并选用合适的评估方法。本文提出的基于 SOA 架构的评估系统框架,以 Web service 形式动态调整评估方法以适应不同的应用领域,重点不在于如何改进和设计评估方法,而是立足于数据的源头,对数据内容及其演化过程的世系技术进行研究,研究其对数据质量评估的支撑。其难点在于不同环境下数据维度的确定,因此我们考虑引入数据世系技术,重点研究数据内容及其演化过程,设计了世系的语义模型,为世系数据的进一步操作奠定了理论基础,同时对世系数据的查询给出了一个理论模型。但是由于世系数据比较复杂,不仅需要查询其内容,在对数据质量进行评估中还需要了解其演化过程,因此该查询模型还需要进一步完善,同样由于篇幅问题对于世系数据的采集等内容没有在此进行探讨。今后的工作将继续完善该框架的各个细节,包括基于结构的世系查询模型、世系数据的自动采集和存储、如何通过世系数据确定评估维度等。

(下转第 492 页)

成，极大地提高了各种操作的运行效率和速度，同时也很好地屏蔽了数据表的逻辑结构和数据表之间的关系。公共操作层与业务无关，但是被经常使用的功能抽取出来，作为独立的基础服务模块。这些模块具有较强的可重用性。包括一些文件操作、邮件发送等相关的函数。

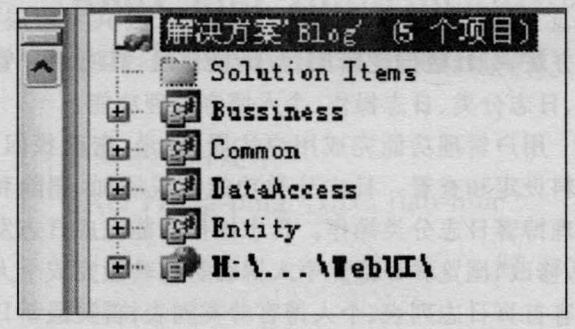

图2 多层Blog系统的项目结构

3 结束语

基于ASP.NET +SQL Server的多层Blog系统采用多层体系架构，将系统分为公共实体层，业务逻辑层，数据逻辑层，数据操作层和用户界面UI层。多层Blog系统结构清晰、各层功能明确、易于扩展，如果增加业务功能，只需在业务服务层增加功能模块即可，能有效降低耦合，有利于程序调试和开发。多层Blog系统的应用服务器层根据功能分成若干个层，并利用单独的项目进行管理，极大地提高了系统的稳定性、可扩展性和移植性。

参考文献：

[1] Microsoft Corporation. Designing distributed applications[DB]. MSDN Library For Visual Studio, NET, 2005: 212-214.

[2] 梅宏，申峻嵘. 软件体系结构研究进展[J]. 软件学报，2006,17(6): 1257-1275.

[3] 刘军，阳小华，杨星. 教学信息发布与管理系统的设计与实现：基于.NET组件技术[J]. 计算机工程与应用，2006(2): 99-02.

[4] 郭磐君，张艳丽，王芳芳. ASP. NET AJAX入门与案例精解[M]. 北京：机械工业出版社，2007.

（责任编辑：邓大玉）

（上接第485页）

参考文献：

[1] Wang R Y, Trong D M S. Beyond accuracy: what data quality means to data consumers[J]. Journal of Management Information Systems, 1996, 12(4): 5-33.

[2] Pipino L, Lee Y, Wang R Y. Data quality assessment [J]. Communications of the ACM, 2002, 45(5): 211-218.

[3] 韩京宇，徐立臻，董逸生. 数据质量研究综述[J]. 计算机科学，2008, 135 (12): 1-5, 12.

[4] Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement[J]. ACM Computing Surveys, 2009, 41(3): 1-52.

[5] 杨青云，赵培英，杨冬青，等. 数据质量评估方法研究[J]. 计算机工程与应用，2004. 9: 3-4, 15.

[6] 高明，金澈清，王晓玲，等. 数据世系管理技术研究综述[J]. 计算机学报，2010, 33(3): 373-389.

（责任编辑：韦廷宗）